基于媒体语料库的检校体系探索

高 文
（新华通讯社通信技术局，北京 100803）

摘 要：新闻采编中除一些印刷错误外，很多错误是潜在的语义级错误。语义错误需要检查语句中所表达的语义和语用是否违背了某种标准，采用以往的文本检校方法，很难发现这些语义错误。例如，报刊、网络文章中出现的一些关于中国台湾问题的不正确表述等，利用自动检校工具来纠错难度是相当大的。但类似错误对新闻机构的影响不可小觑，部分可能是影响舆论导向的政治性错误，是编辑部检校的重中之重。因此，充分利用媒体语料库的新闻采编检校是新闻文本自动检校的重要发展方向。

关键词：新闻采编；检校；语料库；机器学习；知识图谱　　**中图分类号：**H131.2　　**文献标识码：**A
文章编号：1671-0134（2021）11-142-03　　DOI：10.19483/j.cnki.11-4653/n.2021.11.044
本文著录格式：高文 . 基于媒体语料库的检校体系探索 [J]. 中国传媒科技，2021（11）：142-144.

在中文检校领域，目前市场上有很多检校软件和服务，在语料库覆盖、词库容量、检校算法、系统兼容等方面调研的结果分析和实际使用体验表明，由于软件本身的局限性，在语义层面的检校能力有限，部分词库不能自动在线更新也是一大弊病，仍需要检校人员做进一步的人工检校。然而采用人工检校，劳动强度大、成本高，并且由于检校人员的责任心或视觉疲劳等问题，仍会漏掉一些错误。有时，某些错误，尤其是带有语义的错误是会被放大的，甚至被某些居心叵测的人或媒体利用，造成不良影响。这就进一步要求在检校技术中侧重研究含语义分析乃至知识库先验策略的检校技术，以弥补现阶段检校软件的缺陷。

1. 利用新华社现有媒体语料库构建有效元数据

语料库是应用计算机技术对海量自然语言材料进行统计分析的大型资料库。新华社作为国家通讯社，具有海量且权威的新闻语料库，对构建新华社新闻报道专属媒体语料元数据具有绝对优势。典型的语料库系统应包括：文档的抽取及元数据创建；自动词性 / 语法标注；索引、检索和统计分析等功能模块。其中最为重要的是词性、句法、语误的标注环节，系统提供的标注手段和准确率直接关系着语料库的建设规模大小和研究成果的优劣。[1] 故而，在充分利用新华社多媒体数据库中丰富的新闻语料构建专用元数据的同时，更需要经验丰富的编辑和检校人员配合筛选和提供用于训练检校模型的正例和负例样本，这些样本集的先验数据的可用性将会直接影响检校模型的检错率。

机器学习和深度学习算法可以从数据中自动挖掘统计规律，学出计算模型。[2] 数据的质量和数量对模型的效果至关重要，近年来深度学习获得成功的一个重要因素便是大数据的支持。对新闻采编的语义校对来说，建立大型专属语料库必不可少。

1.1 已发布的新闻内容数据

新华社在长期的新闻采编工作中积累了海量且权威的新闻语料，对构建新华社新闻报道专属媒体语料库具有绝对优势。这些语料经过了严格的软件和人工校对，具备高度的准确性和规范性。同时，其语言特性，词汇和短语搭配的统计特性，可以为模型训练提供丰富的正样本。该部分语料的规模将远超以往学界研究中所用的规模，为训练大规模深度学习模型提供有力的数据保障。

1.2 经过专家修改的新闻文本数据以及修改前的数据

负样本对机器学习模型是必不可少的。学习算法需要从负样本中挖掘语义错误的各种模式，并进行推广。负样本中的语义错误应当和实际中的错误有尽可能相似的特性。新华社的校对人员在其长期工作中，检查并改正了各种类型的语义错误，保障了新闻报道的准确性。这些人工校对检查出的错误，以及修改前的数据将是非常有价值的负样本数据，并为负样本的自动生成算法提供非常好的参考和启发。比如，有些语义错误往往是输入了错误的同音或近音词，该词本身没有错误，但是和前后词语搭配起来在语义上明显不同。针对这个规律便可以对正样本进行随机同音或近音词替换来生成负样本。通过采集和生成这两种手段，可以获得海量且高质量的负样本数据。

1.3 大量的全网新闻文本数据

除了新华社提供的权威数据，对互联网上可以搜索到的大型纸媒或者网络媒体，也存在大量的新闻数据可以作为训练样本。通常采用数据抓取技术，对这部分数据进行抓取，形成更加全面和强大的新闻数据库。目前网络爬虫技术已经非常成熟，例如今日头条、百度新闻中都对媒体的数据进行了抓取和聚合。在爬虫技术中，加入正文提取技术进行辅助，因为对不同数据源的数据，网页结构都有不同，故而使用更具针对性的爬虫设计，

可以实现对不同网站新闻数据的海量爬取。

1.4 领域知识数据

新闻采编中可能会遇到一些知识性错误，比如人物、事件、地点和时间等发生了错误的对应和搭配。这类错误涉及的语法和前后语义搭配没有问题，仅仅依靠语言本身的特性无法成功检测，必须利用和报道领域相关的专有知识。知识图谱是一种对信息进行结构化组织和表示的工具。它将各类事物表示为实体，将事物之间的关系表示为实体之间的各种链接，并用这种方式将人类世界中的知识组织起来，从而为各种与知识相关的应用提供支持。知识图谱在搜索引擎、信息检索和自动问答中已有非常成功的应用，多年以来一直是各大科技公司和学校研究的热点，相关的技术也比较成熟。

在检校体系构建中，此部分包括知识图谱的构建和应用两个阶段。第一阶段是建立针对新闻报道的大规模知识图谱。这部分工作通过多种算法的组合来实现知识图谱的自动构建。首先是对已有大规模知识库，如百度百科、维基百科进行知识提取，其中包括大量结构化知识，仅需要简单的操作便可高效转化到知识图谱中。此外，互联网网页中也蕴藏了海量的知识，这些知识往往以非结构化的形式存在，比知识库要杂乱一些，但通过自然语言处理等自动化技术也可以将其抽取出来，并加入知识图谱。最后，对多个来源的知识进行融合，对知识图谱内的实体关系进行推理，剔除一些错误的知识，进一步改进知识图谱的质量。第二个阶段，借助知识图谱设计语料知识的查询和验证算法，实现对知识性错误的检测功能。包括新闻语句中的实体抽取，实体链接，与知识图谱中相应的内容进行对比并返回结果等。

2. 机器学习辅助实现自动新闻文本纠错模型

语义分析中可使用条件随机场，深度学习卷积神经网络、知识图谱等机器学习模型实现自动新闻文本纠错模型的构建。

综合使用这三种先进的算法将可实现更加优化的新闻文本校对算法。语义错误的检测依赖于文本的上下文信息，这就要求模型能够学习到一定范围内的上下文的依赖关系，该范围可从短语到单条语句甚至多条语句。

2.1 概率模型纠错预测

概率图是将图结构和概率统计相结合的一种模型，适用于针对特定结构的数据的推理和预测任务，在文本这种序列数据的处理中有广泛的应用。条件随机场是一种处理上下文信息常用的概率图模型。条件随机场从多个连续字词中抽取特征作为输入，并计算输出标记序列的联合概率分布，输出标记可以根据应用的目标而定。[3] 输出标记可以对应词语的适合程度，当适合程度过低时将其归为错误。条件随机场在多个自然语言处理任务中都有成功的应用，包括词性标注、句法分析和命名实体识别等。输入特征可以使用词性特征和词向量特征等浅层特征，这些特征在工业界中已有广泛应用，并取得了不错的效果。条件随机场可以学习到前后词语搭配组合的统计特性，并据此对语义错误进行预测。

2.2 深度学习建模

条件随机场是一种传统的浅层模型，模型的复杂程度较低，只能学习到相对较小的上下文依赖关系。近年来深度学习模型在人工智能的各类应用中都取得了优异的效果，它通过多层神经网络从数据中学习高度抽象的特征和统计规律。由于模型的建模能力和计算的并行程度都很高，深度学习非常适合与大数据结合的应用，能够充分地利用海量语料库的优势，学习词语搭配中较长范围的上下文依赖关系。

使用卷积神经网络和递归神经网络来完成语义校对的功能，首先对语句进行分词的预处理，每个词用初始化的词向量表示，然后将词向量的长序列输入到深层神经网络中，最终输出每个词存在错误的概率。[4] 整个过程除了分词的预处理步骤，完全是自动学习的，不需要任何手工设计的特征。

2.3 基于知识图谱进行更高级的语义纠错

深度学习可以学习较大范围的上下文依赖，但是很难对知识性的错误进行检测和纠正。这类错误的校对需要依赖知识图谱来完成。这部分任务的主要内容是设计实体抽取、实体链接和实体关系抽取等算法。实体抽取是从语句中抽取表示具体人物、地点、物品等的实体。一种方法是通过词性标注来找出语句中的名词，它们往往和实体对应。还可以建立专有实体的词库，并对文本进行匹配查询。实体链接是将抽取出的实体和知识图谱中存储的实体进行联系和对应，消除实体的模糊性和可能的歧义。[5] 这部分任务也需要利用上下文信息，可以再次应用前文中所述的条件随机场和深度学习模型。实体关系抽取是从语句中抽取两个实体之间的语义关系。有很多方法可以用于关系抽取，包括基于特征的监督学习方法，基于自展法的半监督学习方法，基于聚类的无监督抽取方法。完成这些步骤后，便可以在知识图谱中查询和比对，并返回知识的验证结果。

3. 基于媒体语料库的检校软件研发的切实意义

充分结合新华社既有语料库，规范新闻采编用语。新华社每天签发 2000 余篇稿件，其中仅中文稿日均稿量也有近 800 篇。除此之外，新华社还发布了《新华社新闻报道中的禁用词》等，对若干领域的新闻报道用词加以规范。这些丰富且专业的语料对新华社而言无疑是宝贵的财富。同时，深度学习的先决条件就是需要大量的有效数据。因此，新华社的新闻语料库对检校模型的训练具有重大意义。以语料库中已有的正确数据作为先验，结合生成或添加的大量负例样本，迭代完善检校模型，从而实现对新闻报道用语规范的监督校验作用。

优化检校流程，提高工作效率。目前各大编辑部使

用的检校软件能够解决部分别字类错误，加入语义级错误的识别功能无疑是对检校流程的进一步优化和改善，提升新闻工作效率，同时也更好地保证新闻产品的产出质量。

值得提出的是，自动检校工具在新闻工作中仅能作为辅助工作的手段，并无法完全避免所有文字和语法语义上的差错。因此，自动检校工具软件在优化工作流程和提高工作效率的同时，为新闻工作者节省一定的时间和精力，并不意味着可以省去人工审稿的环节，而是有更多的精力去优化新闻稿的采写。

训练专有检校模型，更好地服务于不同内容领域的检校工作。在不同新闻内容领域，如教育、医疗、能源、政治等，常用词和专有词汇交集很少，在专有领域利用深度神经网络训练特有的检校模型对该领域的语义级识别将更准确。目前新华社使用的编辑系统中，每一篇稿件都具有稿件分类的专用字段，利用该字段对检校模型进行选择，调用相应检校算法对当前文稿进行分析，获得更准确的检校结果。

结语

基于媒体语料库的智能检校并不是一个简单的技术任务，需经验丰富的新闻工作者的支持。基于媒体语料库的深度学习，毋庸置疑，语料库的选择是决定性条件之一。在新华社多媒体数据库中，有海量的新闻稿作为深度学习的正例，负例的构建或生成就需要借助经验丰富的新闻工作者的积累，对以往典型的、常见类型的错误进行收集作为负例中的重要部分。深度神经网络的多层结构的好处就是可以用较少的参数来表示复杂的函数。同时，这又要求训练样本数据能够尽量覆盖未来的样本，那么学习到的多层权重便可以很好地用来预测新的样本。故而，样本的构建需有丰富编写和检校经验的新闻工作者的支持，筛选和构造正负例样本，以及不同别字和语义错误的出错率权重的判定。只有得到足够多的有效样本，对未来测试样本的识别率才可能得以提升。新闻工作者在稿件采写和检校的丰富经验与人工智能技术紧密结合，才能使工具软件为新闻工作提供更好的服务。媒

参考文献

[1] 卫乃兴 . 基于语料库和语料库驱动的词语搭配研究 [J]. 当代语言学，2002（2）：101-114，157.

[2] 王文通，王立春 . 深度学习研究综述 [J]. 北京工业大学学报，2015（1）：48-59.

[3] 洪铭材，张阔，唐杰，李涓子 . 基于条件随机场（CRFs）的中文词性标注方法 [J]. 计算机科学，2006（10）：148-151，155.

[4] 周飞燕，金林鹏，董军 . 卷积神经网络研究综述 [J]. 计算机学报，2017（6）：1229-1251.

[5] 刘峤，李杨，段宏，刘瑶，秦志光 . 知识图谱构建技术综述 [J]. 计算机研究与发展，2016（3）：582-600.

作者简介：高文（1985-），女，山东东营，工程师，研究方向：计算机应用技术。

（责任编辑：张晓婧）

（上接第135页）

坚定道路自信的集中体现。如今的江教集团已然探索出一条适合自身发展、符合国家要求、响应人民期待发展道路，正成为江西省内，乃至全国教育传媒业转型创新中的典型样板。与市场化、企业化转型相适应，江教集团迅速解放思想，建立起相匹配的人员选拔机制、考核激励机制、质量管理机制等与传媒企业相适应的现代企业管理制度，通过机制的创新驱动事业的发展，江教集团的综合实力进入全国30余家省级教育报刊社第一方阵。

结语

借用习近平总书记在全国教育大会上关于“坚持以人民为中心发展教育，核心就是办好人民满意的教育”的观点，那么做好教育服务与宣传，其核心就是加快构建以人民为中心的教育传媒发展格局。坚定不移遵循服务群众、依靠群众、联系群众的根本原则，继续深化教育传媒改革，加强宣传与引导，不断满足人民对教育信息服务的多样化需求，真正让人民共享教育传媒发展的新成果。媒

参考文献

[1][美] 保罗·拉扎斯菲尔德、罗伯特·默顿 . 大众传播的社会作用 [M]. 北京：人民日报出版社，1983：169-170.

[2] 方玉 . 新时代下教育新媒体编辑素养的提升路径探索 [J]. 新闻研究导刊，2020（6）：130-131+208.

[3] 梅宁华、支庭荣 . 媒体融合蓝皮书：中国媒体融合发展报告（2016）[M]. 北京：社会科学文献出版社，2017：2.

[4] 姜圣瑜 . 从“受众时代”走向“用户时代”[J]. 传媒观察，2011（4）：24-26.

作者简介：方玉（1976-），女，江苏阜宁，副编审，研究方向：教育新媒体、教育媒体融合。

（责任编辑：张晓婧）